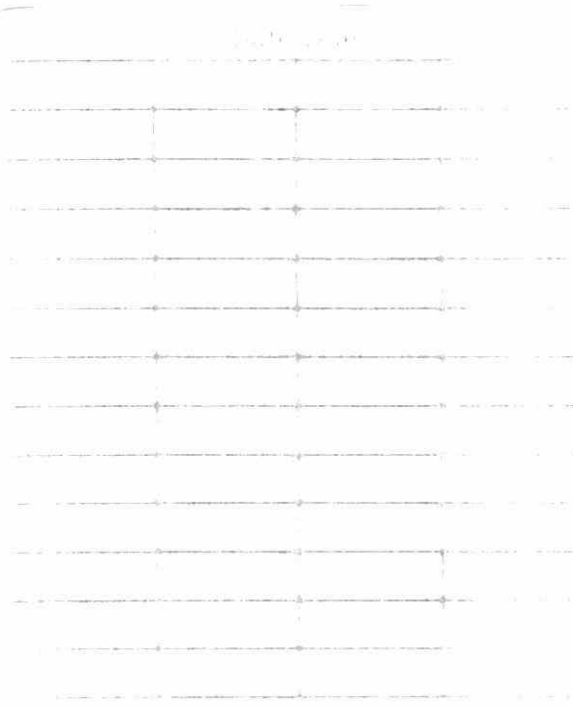
DATA QUALITY REPORT SERIES

CREDIBILITY: THE CONSEQUENCE OF QUALITY ASSURANCE

Ministry of the Environment
rio

The Honourable
Keith C. Norton, Q.C.,
Minister

Gérard J. M. Raymond
Deputy Minister



CREDIBILITY:
THE CONSEQUENCE OF
QUALITY ASSURANCE

BY

DONALD E. KING, M.Sc.
QUALITY ASSURANCE OFFICER
LABORATORY SERVICES BRANCH
ONTARIO MINISTRY OF THE ENVIRONMENT

MAY 1982

Credibility: The Consequence of Quality Assurance

Summary

This paper reviews the relationships between quality control, quality assurance and credibility. Various factors important to the establishment of data credibility are enumerated with particular emphasis on those which can be readily addressed within the laboratory environment. Quality assurance is described as a management function which rests on the documentation and establishment of quality control protocols, and on the evaluation and summarization of their outcomes. Quality control is a technical, operational function which investigates and confirms the proper conduct of all those procedural components necessary to a successful conclusion. The outcome of quality control operations in the laboratory allow one to determine the repeatability, reproducibility and bias of the analytical system, in particular the analyst, as well as to establish the precision and accuracy of the analytical procedure, and the stability of the instrumentation used. Techniques for establishing bias and error are briefly outlined. The terminology involved is reviewed. Finally there is a brief discussion of the problems of reporting data, particularly at or below the detection limit. The difference between a detection criterion, and the detection limit is reviewed.

Credibility requires action by more than just the laboratory analyst. If the result does not reflect the real world because of inadequate program planning, field operations or inappropriate data use and interpretation, the promulgation of laboratory quality assurance documentation can work against credibility. However, the laboratory analyst is central to the entire operation, since he can see the results of inappropriate activity by his associates and is best able to advise them on remedial action.

Credibility: The Consequence of Quality Assurance

Introduction

Credibility is fast becoming the catch-word of the 80's for those involved in the study of environmental and health effects arising out of the treatment and disposal of waste material. In the face of contradictory evidence from analysts supporting opposing view-points, the public, and even data users within the same agency as that providing the analytical service, are left uncertain as to whom to believe. Expressions of confidence from the analyst based on even the best in-house control activity is increasingly being perceived as insufficient or at worst as a whitewash. One needs only to become involved in the evaluation of interlaboratory comparison data to realize, in spite of the emphasis placed on quality control during the 70's, that something significant is missing.

Credibility lies in the eye of the beholder. There is always a strong tendency to attach credibility to facts which support one's own particular beliefs. (All other facts are incredible!) Analysts, like anyone else, believe strongly in their own credibility because facts, to which they alone are privy, support the viewpoint that what they do produces reproducible data. When faced with a challenge from another analyst, they know the degree of credibility of their own operation, and automatically assume that the other fellow's results are very nice but wrong. Of course the feeling is mutual. No matter how much time one spends on establishing one's own confidence, there is no guarantee that anyone ese will be convinced, unless there are well documented facts to back it up.

Establishing Credibility

The subjective and volatile nature of credibility demands that it be established in an organized logical fashion, and that the facts supporting it be well documented. Each facet of the operation must be examined in minute detail, with particular regard given to those which are most subject to attack. In the case of environmental investigations several factors must be addressed. In broad terms these include;

a) the rationale for the investigation. (What are we trying to prove and, equally important, disprove?)

b) definition of resource needs relative to existing capability. (Will we be able to find what we are looking for?)

c) the design of the field sampling program. (Will it produce all data required, both in support of and against taking the action proposed by the topic under investigation?)

d) the field activity necessary to obtain the samples in the form required for analytical evaluation. (Will the samples be worth analyzing?)

e) the maintenance of sample integrity and identity. (Are samples properly and sufficiently identified? Will the reported data actually match the right samples?)

f) the proper interpretation of data. (Will our conclusions be supported by the data presented? Will all data be evaluated?)

The above factors tend to be beyond the scope and, to some extent, beyond the influence of the analytical scientist in most large agency laboratories. But there is growing concern that unless they are properly looked after, the analyst will be left holding the bag. It is difficult to establish data credibility when one is unable to establish the credibility of the sample or the sampling process, the chain of custody or suitability of transportation facilities.

While we realize that the larger issue of credibility of an analyst with the general public mainly falls outside the immediate control of a laboratory quality assurance program, it is apparent that any data user would be extremely foolish to ignore the consequences of lack of public acceptance of his data. As stated above, many of the factors influencing the analytical result are beyond the control of the analyst. To this extent therefore, the analyst cannot be held responsible if the sample which he examines does not truly represent the situation under study. Under such circumstances, however, where the analytical result does not then reflect the real world situation, the promulgation of QA documentation becomes a cruel joke on the organization concerned and can actually work to destroy credibility.

It would seem, therefore, that some power beyond that of the laboratory analyst is necessary to establish and maintain the public credibility of scientific organizations. We can all cite examples of organizations whose scientific pronouncements are seldom questioned. Examination of these situations generally reveals that the spokesperson for the group is a charismatic individual of demonstrated sagacity. Closer examination of his support group may reveal dedicated commitment to the aims of the organization or it may reveal sloppy methodology leading to highly dubious analytical results. In either case, however, the organization makes progress by demonstrating an ability to use the results in a positive fashion for the publicly perceived betterment of a situation.

In view of the fact that charismatic leaders are in short supply these days, what can be done to enhance the overall acceptance of our data? I am afraid that there is no simple answer to this question. There is, however, a philosophic approach which when assiduously pursued minimizes those occasions when published data assails the public credulity.

The first step is to apply the test of reason to each analytical result, i.e. could the sample, as described, contain this substance in this concentration? The second step is to question a reasonable result, i.e. is this the most likely number to apply to that substance in this sample? The third step is to assess the acceptability of the result to the client, i.e. is this the result which will be most gratifying to the submitting agency? The fourth step is to prepare an appropriate dialogue for use when the answer to step 3 is negative, perhaps including a written comment on the report to explain the result.

In summary the same painstaking care with which the result is determined should be applied to ensure that the result is appropriate for the intended use. The end product of laboratory analysis should not be numbers, but solutions to the client or data user's problems.

Credibility and the Laboratory

Within the laboratory there are of course more than enough areas to be addressed in order to establish credibility. Factors to be considered include;

a) the physical structure and condition of the laboratory facility. (Is the lab clean and tidy and properly laid out for the safe performance of analytical functions?)

b) the mechanical condition of equipment used for analysis. (Is it properly maintained and correctly, and safely, connected to all power, water, gas, etc. sources required?)

c) the source and quality of analytical reagents, distilled water or other solvents, and their storage. (Are precautions in place both to protect the user and to ensure reagent integrity?)

d) the availability of completely documented analytical methodologies appropriate to the sample types being examined. (Are they taken from standard reference texts?)

e) the existence of documented quality control and quality assurance protocols

f) the proficiency of the analyst, and his/her experience with the sample types and analytical procedures routinely encountered. (Is she/he trained to recognize and handle special cases?)

g) the source, reliability, preparation and maintenance of calibration materials, in-house controls and external reference material. (Can someone else backup your values?)

Most of these factors are defined and recorded at a given point in time. They should be reviewed periodically to ensure the necessary standard is being maintained. Procedures for inspection, verification and documentation of current status must therefore be developed. In fact it is this activity that involves the concept of a laboratory quality assurance program.

Quality Assurance

The rationale behind establishing a laboratory quality assurance program is to provide a basis for defining and documenting lab and data credibility in terms

acceptable to both the individual client and the public-at-large. It is completely immaterial that an analyst performs any of a number of controlled operations in order to ensure that he can confidently attach his name and reputation to an analytical report, if, in the face of a challenge from another analyst or the opposing side, he cannot provide adequate proof that these operations were carried out, that their outcomes were reviewed and evaluated, and that continuing control was demonstrated.

Many laboratory managers and analysts do not yet appreciate fully the difference between quality control and quality assurance. Because they are exposed at first hand to the data generated by their own control operations, they develop a "feel" for their data quality. They become "confident" because the numbers generated or the observations made in order to monitor the variation of a particular component in their analytical process, are repetitively contained within a narrow range. When pressed they can casually refer to their past experience with respect to the "precision" and "accuracy" of their data. However if pressed further they are often forced to admit that raw data required to substantiate their claim is not readily available. Even when it is in a more or less presentable format, the fact that another confident analyst has produced results which appear to be in direct conflict, makes it all too clear that part of the story must be missing.

Quality assurance therefore rests on one's ability to retrieve documents:

a) which establish the daily protocols implemented to monitor the various factors which have been known, or are most likely, to affect the total measurement system, and specify what action is to be taken if trouble occurs.

b) which review the quality assessment techniques employed to evaluate data produced in support of the above monitoring protocol, indicating their frequency of application.

c) which identify supervisory responsibilities and duties in order to maintain quality assurance.

d) which summarize and review the degree of success achieved in daily operation, i.e. the limits within which the system was actually maintained,

the number of occasions when it failed or almost failed, and the significance of such for the various data user groups being served.

e) which provide evidence summarizing the comparability of data, i.e. against what standards, or in association with which other analysts, and with what frequency and success has comparability been established.

f) which review the historical development and implementation of the current procedures, i.e. against what other methods were they compared, what range of sample types, and how many, were used to establish the suitability of the method, and how does current data compare to that provided under the previous methodologies.

It should be apparent that quality assurance is a middle-management function. Supervision, evaluation and review, summarization and documentation, are the key words. These do not represent tasks generally assigned to a junior technician, or a scientist working on the bench. It is only recently becoming apparent that quality assurance protocol for supervisors and managers has not been given the attention it requires. There is no question that familiarity with the usual procedures, and the general routine tends to preclude any urgency in documenting them or their outcome. Scientific staff in such positions must give consideration to documenting how their QC protocol works, the rationale behind it, and the expected impact on data and laboratory credibility in order to ensure that their staff are knowledgeable and "on track". In this regard, the quality assurance oficer, if such a position exists, provides guidance and assistance in this documentation effort, but the effort must originate with, and reflect, actual lab management and supervisory activity.

Quality Control

Quality control activity provides the data upon which quality assurance reports are based. There is no question that a technician can maintain an analytical process in control by carefully checking each of the many factors involved. This may or may not require a tabulation of the specific observations. However "control" implies preventive maintenance rather than emergency response. Prevention implies foresight and a predictive capability. Therefore true "control" requires firstly, on the spot tabulation or charting of the observations required by the established protocol, and secondly, documentation of action taken if required.

Control charting allows one to observe trends toward an out-of-control situation as well as pick out specific instances of loss of control. It has been noted that:

a) a system must be established to be in control, in order to be maintained in control.

b) a system is not in control if it is observed to produce unexpected data more than once every twenty to twenty-five runs.

c) control limits usually become tighter once a process is placed under a control protocol as a reflection of the fact that the original limits set were based on data produced by an uncontrolled operation.

d) if the process, under control, produces information which is better than that determined as essential for the data use intended, the establishment of acceptance limits looser than the control limits is permissible. The two modes of operation should not be confused however. (In some systems, day to day variation is more difficult to control than within-day variability would suggest. If one loosens the within-day control process it becomes even more difficult to detect the between-day changes which are potentially more serious since they are related to maintenance of accuracy rather than simply precision.)

Quality control should be present from start to finish. It not only provides evidence of product quality, but it also ensures that product quality will be maintained at a high level by providing a check on the components that go into it. As noted earlier a system is not "in control" unless it has been consistently within the control limits over an extended period of time. If an analytical run is set up every working day, failure to meet the control limits more than once a month suggests some component of the process is out-of-control.

Component control, then, is the preventive action necessary to maintain product control. It includes regular checks on analyst proficiency, method ruggedness, reagent quality, media preparation, solvent and distilled water purity, laboratory glassware, cleanliness and storage. It also monitors equipment operation in the form of maintenance schedules, etc. for replacement of parts before they wear

out, and in the form of charts for temperature, humidity stability, etc., where important. The preparation of analytical calibration standards should be controlled to ensure they agree with previous standards or in-house controls before use. Equipment response should be regularly checked against appropriate standard materials i.e. thermometers, balances, spectrophotometric wavelengths, GC retention times for freshly packed columns, etc.

All observations made for component control should be recorded in a permanent, organized fashion for ease of retrieval and review. Control charts may be necessary where the ruggedness of the system depends on tight control over that particular factor. This recording or charting activity does not of itself provide quality assurance, but does provide for easier supervision to ensure protocol is being maintained and appropriate action is being taken. Product control is generally assured if component control has been properly maintained. The truth of this must, however, be documented in order to determine the precision and accuracy of the analytical process is being maintained.

Method Documentation

Laboratory credibility is generated as a result of the existence of documented protocols which specify the tasks undertaken to ensure that an analytical result is properly measured and correctly interpreted. A carefully described analytical method is critical to the analytical outcome. Traditionally analysts have depended upon the texts of "standard methods" promulgated by various standards-setting organizations such as APHA, ASTM and US-EPA, but the explosion in technological progress since the 1950's which has seen the development of automated, microprocessor-controlled instrumentation in a variety of applications, has made it difficult for such texts to keep up-to-date. Increasingly the instrument manufacturers have provided new detection systems based on physical rather than chemical processes, and have incorporated into their systems the calibration and calculation functions. Under these conditions the nature of the methodology document, and the quality control functions which must be an integral part of the method, must change to include the specific instrumentation employed, in relation to the extent to which it has displaced the analyst. We must now control the instrument as well as the analyst. Methods must be described relative to the time and place and conditions in which they are employed. Traditional "standard methods" do not

adequately address this need, therefore the analyst must provide his own documented procedure, incorporating all those steps required for sample preparation, instrument and reagent and standards preparation, as well as the quality control checks performed along the way.

Quality Control Terminology

Control is exerted in order to prevent error, not to monitor it after it has occurred. There are two types of control; component control investigates the suitability of the ingredients and the correct operation of all necessary equipment (including the human), whereas product control establishes the quality of the data produced. "Standard methods" are accepted and defined as such partially on their "ruggedness", i.e. non-susceptibility to small changes in procedure. Therefore some factors need less control than others. A particularly critical factor, however, is and will always be the human operator.

Product control is, in fact, a measure of the performance of the chemical/physical, instrumental, operator team. In general terms we wish to observe and thus control the "precision and accuracy" of this team. In order to do this effectively we must break the total down into manageable pieces. This requires careful definition of terms. The following are recommended because they provide clarification of the distinction between terms often considered to be synonymous.

Precision - an inherent property of the method, which a very proficient operator may ultimately achieve. Basically it is the ability of the method to produce a very tight range of values for repeat analysis of standards.

Accuracy - an inherent property of the method, including the physical-chemical principles involved, which determines how close to "truth" the average of a series of precise measurements will be. Truth is generally defined by international agreement.

Repeatability - a property of the analyst dependent on proficiency and analytical conditions, the nature of the sample and the range of concentration of the analyte. It indicates ability to produce a tight range of values for repeat analyses of samples once the analytical system has been set-up and calibrated.

Reproducibility - a property of the analyst and the stability of the "instrumentation" employed. It indicates ability to achieve essentially the same calibration from one day to another (or from one analyst to another). It should be not more than 50% larger than the repeatability if proper calibration control is established.

Repeatability and reproducibility are estimated respectively by, the within-run standard deviation (S_w) obtained usually from duplicates analyzed within the same calibrated run, and the between-run standard deviation (S) obtained usually from one or more control samples analyzed on a day-to-day basis. The use of paired controls permits estimation of both S_w and S and thence identification of systematic error which is the source of non-reproducibility.[2]

Deviation - the natural random variability introduced in any measurement process due to indeterminate changes from the exact procedure.

Error - deviations which are so large that one must conclude that the exact procedure has not been followed correctly. The source of the error may not be apparent. Obviously a reference point is required, either a previous average value or a single other result. In the latter case error is suspected rather than confirmed. It may result from randomly contaminated glassware, etc.

The terms Indeterminate and Determinate Error are not needed if the above discrimination between Deviation and Error is recognized. This is also consistent with the use of the term Standard Deviation as a measure of range of Deviation.

Systematic Error - error introduced (most typically) in the calibration process. It will affect all data produced in a given run in the same way. An intercept error has a constant effect whereas a slope error shows a proportional effect dependent on concentration.

Bias - difference on average between results produced by different analysts and/or different methods. The difference may be constant if an intercept, blank or background estimate differs between the analysts etc., or it may be propor

tional to concentration if the standards used for calibration do not agree. A bias of less than 10% is to be considered good. Bias at or near zero is of greater concern, and should be stated in absolute (i.e. mg/L) rather than relative (i.e. %) terms.

Inaccuracy - difference between an in-house set of standards and an external reference standard. (In order to be confident that a difference equivalent to one standard deviation will be detected requires at least 13 replicate analyses of both the in-house and the reference materials.)

In general, imprecise data cannot be used to determine the degree of inaccuracy in absolute terms. We can only say that the magnitude of any inaccuracy will not exceed some value based on the imprecision of the method and the number of replicate analyses performed, neglecting for the moment any consideration of bias.

Specificity - a characteristic of the method and its ability to isolate and detect a specific target element or compound. Many tests are "specific" only under the assumption that other interfering constituents will not usually be present in the routine samples for which the method has been developed. The test-method nomenclature should, but does not always, properly identify the target compounds or elements.

Recovery - a factor properly associated with the sample preparation process but generally confused with a correction for the effect of the sample matrix, etc. on the calibration curve. In some instances the recovery factor may be corrected for, but in others, particularly organics, it is often reported separately.

Calibration - the complex protocol of establishing, at a point in time, the relationship of response against concentration. It estabishes the traceability of in-house standards and controls against the values obtained on external reference materials, as well as the linearity and stability of the analytical system.

Standardization - the day to day process of confirming the measurement system is kept in calibration.

Any proper methodology documentation process will address these topics and incorporate specific quality control operations, where necesary, to ensure that data quality can be maintained and demonstrated to exist, in each of these areas of system proficiency. The protocols for gathering, evaluating, and summarizing the outcome of such quality control activity are not incorporated into the methodology. In the same manner as instrument operation, they form part of the training process to which the new analyst is exposed, and will be documented as standard operating principles.

Repeatability vs Reproducibility

For the individual analyst, it is suggested that true precision and accuracy cannot be independently established. Precision must be considered as an attribute of the perfectly operated process rather than of the analyst. Even a perfectly operated system will have an inherent bias or inaccuracy which depends upon the physical or chemical prinicples involved, and which may remain undetermined if an alternative independent system is not available.

Repeatability, on the other hand, is an analyst attribute. It changes in magnitude from one to another, and from time to time, dependent on experience and proficiency. It should be determined/estimated for the analyst, before he is allowed to work alone, based on his ability to obtain satisfactory replicate results from real samples. Usually duplicates are employed. The analysis is repeated on a series of samples, within the same calibrated run, in such a way that the duplicates are non-adjacent and, preferably, in a "blind" location in the run.

Repeatability establishes that a state of "simple statistical control" exists for the measurement process on a within-run basis.[1] However chemical analytical procedures depend on restandardization from run-to-run. Therefore a state of "complex, or multistage, statistical control" must be recognized for between-run data. While it is possible "to act for the moment as if" simple statistical control exists, the same is not valid for complex statistical control unless a calibration control protocol has been established to demonstrate it.

Thus, once within-run repeatability has been established, the analyst's ability to be reproducible from run to run must be established. Reproducibility is largely

dependent on the stability of the calibration process. The validity and variability of the daily analytical calibration step must be confirmed by use of paired, long-term, in-house or external control materials.[2] These two "controls" are checked immediately after the standardization of concentration versus response has been established. The difference of the two results monitors stability of the calibration slope, and its standard deviation can be used to estimate within-run repeatability because any bias in the individual values tends to be cancelled by subtraction. The sum of these results, on the other hand exaggerates any systematic error so that the standard deviation of the sums is larger than that for the differences. A great deal of information about the measurement process can be obtained by reviewing plots of sums and differences over time. Thus, if the difference is generally in control, then excess variability in the sum can be attributed to lack of control over the intercept or blank.

If the in-run repeatability of the analyst is established from duplicates, the standard deviation of the difference between paired calibration controls will be approximately $\sqrt{2}$ larger (because two datum are used). Tolerance limits for the sums can therefore be established from in-run repeatability. It can be shown that, for about 50 pairs of calibration controls, the standard deviation of the sums should not exceed 1.87 times that of the differences, to ensure that between-run reproducibility is (statistically) not significantly larger than in-run repeatability, and that a significant calibration variability will be detected if it exists.[3]

Establishing Bias

The above process is required to ensure that the analyst is in a sufficient state of day-to-day control to make it worth while to participate in interlab or intermethod comparison studies. At this point both reproducibility and bias are controlled. The level of error is as yet unknown.

External standards and reference materials are most useful for establishing the presence of bias. Their use, however, is only of value if the findings can be incorporated into the daily operation. The expense and/or relative difficulty of obtaining these materials usually requires the establishment of in-house controls the concentration of which is traceable to external reference materials by repeated analysis of both, within an analytical run which is under both calibration

and repeatability control. This permits translation of findings from today's run into both past and future runs.

When analyzing external materials, the objective should be to detect, and estimate, the level of bias present. Traditionally they have been used to demonstrate that the bias present is not detectable under the conditions of the experiment. Since these conditions are often unfavourable, (insufficient replicate analysis of both the internal and external materials, and inadequate control over reproducibility) it is fairly unlikely that bias would be detected. It can be estimated[4] that in order to detect a difference equal to one standard deviation with a "power" of 90% at a level of significance of 5% (i.e. $\alpha = .05$ and $\beta = .10$) requires at least 13 replicates of each material. If one wishes to ensure that any bias is negligible (i.e. no greater than one standard deviation) the correction factor must be determined very precisely and accurately. More than 50 replicates of each would be required to detect a difference of one-half a standard deviation.

Split-Sample Comparisons

It is possible to compare labs or methods based on analysis of several samples once each way. Over time (given a reproducibility-controlled system) data will be gathered covering a range of concentration. This is amenable to linear regression, and the slope and intercept estimated will indicate the relative bias. The residual sum of squares can be used to estimate the degree of scatter in the data relative to the line of best fit. This can then be compared to the reproducibility estimates for each analyst or method in order to determine how much of this scatter is explained by analytical variation. The remainder is then due to difference between the subsamples analysed and other sample/time related effects.[5]

The correlation coefficient, the average X, Y values and their standard errors are often quoted as evidence of good agreement between methods or analysts. Experience indicates this to be inappropriate. Firstly, it is frequently the case that the averages are not (statistically) significantly different, and yet the slope is not unity and therefore the intercept is not zero. Secondly, the correlation coefficient (r) is only meaningful when it is close to zero. (When comparing methods where the repeatability is small numerically relative to the range of concentration covered, r is always close to unity and tends to depend mostly on the

range and not on the amount of data or degree of fit.) Thirdly, statistical evaluation may indicate a significant difference exists which is so small that it would be essentially impossible to eliminate because of the nature of calibration reproducibility. An average bias between labs over time of less than 10% is fairly acceptable. It is very difficult to maintain bias of less than 5%.

It is interesting to note that if the difference between results is plotted versus the amount reported by one analyst, the correlation coefficient is close to zero. Usually this slope must exceed $\pm$ 0.05 before the correlation coefficient becomes large enough to suggest a difference in "recovery" between the two participants. (This is equivalent to a slope of from 0.95 to 1.05 in the Y vs X plot.) Difference plots are much more useful for evaluating the difference between methods or analysts. If r is small there is no slope difference. It is then and only then that a comparison of the averages is meaningful.[5]

Interlaboratory Studies

Interlaboratory comparisons involve many analysts in an effort to determine "accuracy by consensus". The average or the median value obtained for a given sample is used to set a reference point for identifying anomalous data. One-sample studies are essentially useless except for identifying the very worst performers. Two-sample studies promoted by Youden[6] are better in that the position of points on an X,Y plot identifies the systematic nature of bias between analysts. (If one result is low the other one will probably be low as well.) However the nature of the bias is not identified. It could be proportional to concentration (slope-related bias) or independant of the concentration (blank/intercept bias).

It is often stated that all analysts involved in such studies should employ essentially the same analytical procedure. Experience with a third type of interlab comparison suggests this is not necessary provided sufficient samples and analysts are involved. The bias introduced by a particular operator/laboratory is essentially independent of the method. (If the standards are out 10% it does not matter which method is used.) The use of several samples allows one to determine such things as average rank, following the procedure of Youden[7] and to flag data based on how far it is from the median. The criteria for flagging are set to allow some participants to escape being flagged.[8] These then are identified by their peers as

"very competent". After similar performance on several studies such participants can gain the status of "reference" labs. This type of study helps identify the current state-of-the-art performance potential and often demonstrates that different methods do produce identical data in the hands of competent analysts who otherwise have never had an opportunity to meet or compare notes.

If one wishes, it is then possible to use difference regression to evaluate the difference between each result and the sample median, for each participant. The more competent analysts, given methods with adequate sensitivity and repeatability, produce extremely well defined lines indicating the slope and intercept biases present in their data on the day of analysis. Estimates of their in-run repeatability based on deviation about this line can also be obtained.[5] Under such evaluation one can observe individual points which don't fit the pattern of bias in the rest of the data. Thus an analyst might be biased high by 10% but report one result as within 1% of the median. For that analyst this point although correct is suspicious. By evaluating the distribution of these difference-regression equations it is also possible to observe "clumping" of analysts which may or may not be related to method employed. Thus some may use the same sample preparation but different colorimetric follow-up analysis, and get the same result, or vice versa.

In any event, multi-sample studies tend to reinforce the point that the analyst has either performed very well or that problems exist. Individual points out-of-control for otherwise good analysts, as well as calibration bias, when so clearly identified, require the analyst to take action.

Recovery

One of the findings of inter-analyst study, which arises because of the use of different sample work-up procedures, is that recovery of the constituent of interest is sometimes affected by the procedure used. This cannot be assumed, however, until it is successfully demonstrated that calibration error is not involved. (Multi-sample studies which include a "blind" reference material may provide such proof.)

There is a strong tendency to misinterpret data when it comes to evaluating recovery. "Spiking" the sample with a known quantity of the analyte and then

analyzing both "spiked" and "unspiked" portions is a common approach. The uncertainty attached to the difference between two variable estimates is so great, when it comes to real samples as opposed to clean standard solutions, that whether the calculated recovery is equal to 100% or falls in the range 80% to 120% provides no proof one way or the other of either under or over-recovery. In addition an enormous error in correcting for the "blank" could occur and not be caught. In one study several years ago, three different analysts provided estimates for a drinking water spiked with 50 ug/L of lead of 0.2 and 50.3, 2 and 52, and 100 and 150 ug/L. Obviously all obtained the correct 50 ug/L recovery factor but were variously able to measure what was actually present because of changes in their sensitivity and ability to detect background contamination.

A different technique for correcting for the "recovery" factor will also detect and correct for certain types of error in defining "absolute" zero. It is proposed that it be called "Spiked standard dilution". The sample to be analysed is diluted by a factor of 2 and a factor of 4. These portions, and the original sample, are then each spiked with the same amount of standard, and analysed. The results are obtained after correction for the dilution factor, at which point, if the "spike" was 50 units, there will be apparent spikes of 50, 100 and 200 in the respective results. The results are plotted on the Y-axis versus the apparent "spike" and a line drawn through the points, with greatest emphasis on the lowest point. The true sample result, corrected for both "recovery" and "zero" error, is found at the Y-axis intercept.

This technique is intended for cases where background errors result from contamination of reagents. It will not correct for sample colour or for instrumental background problems. With modification it may be possible to detect distilled water contamination.

Reporting Results

In theory every result reported should include a statement of the method employed, its biases, and effect on recovery. The units of measure should include both the "dimensions" (i.e. mg/L) and the "scale" (i.e. as Si, as SiO_2, etc.). The test name should be non-ambiguous and preferably indicate what was measured rather than what can be inferred. (Some of the traditional names are dangerously misleading,

i.e. Free Ammonia is usually measured as Ammonia plus Ammonium. At one time "Free" meant "distillable", now it is often interpreted as "undissociated" and therefore toxic to fish. Actually, in reasonably hard, i.e. slightly alkaline, water, the porportion of Ammonia to Ammonium is quite small.) There is a move towards the use of Filtered vs Unfiltered, and Total vs Reactive in the nomenclature to clarify the fraction of sample analysed and the severity of the analytical conditions.

The analytical "repeatability" is often quoted, or at least available elsewhere for reference. The quotation of standard deviation should indicate how it was determined. The "reproducibility" is often not known, or is not under control. Statements with respect to "error" should properly identify the magnitude of error that could have been detected if in fact no statistically significant error relative to a reference value was detected.

The statement of a "confidence level" based on multiplying the standard deviation by some factor should best be left to the final user, for the simple reason that there are a multitude of questions that could be asked concerning either an individual result or an assemblage of results. The factor to be used depends on whether it is to be a one-tailed or two-tailed test and whether one requires protection against both type-I and type-II decision errors. Typical questions that could arise include:

a) are these two single estimates different?
b) is this value lower, or higher, than a guideline criterion?
c) what range of concentration could exist in this sample given this result, or average result?
d) is the constituent present in this sample?
e) how much would there have to be in the sample to ensure that it's presence could be reliably detected?

Low-Level Data

Traditionally there has been little consideration given to low results and their correct interpretation. The main question was whether such a "large" amount was present that treatment would be required. Low levels were uninteresting. Nowadays it is recognized that the traditional use of "detection limit" is wrong.[9]

The first problem lies in the definition of "detection limit". It is commonly stated to be either two or three times the standard deviation, but is frequently inflated because of known or suspected error in determining an absolute zero value. In fact the criterion for determining the level above which a result can be taken to mean that the sample probably contains the constituent is 1.64s (s = standard deviation). This is the "Criterion of Detection" with a level of significance of 5%. It is used by the analyst to qualify a result which has been reported. It protects the analyst against Type-I error. On the other hand, the client may use the factor 3.28 times standard deviation to define the point at which a reported result is large enough that a result lower than 1.64s would not likely be obtained by the analyst on a subsequent reanalysis. This represents the "Analytical Detection Limit" with a level of significance of 5% and a "power" of 95%. It protects the analyst against both Type-I and Type-II decision errors, (i.e. concluding the constituent was present when in fact it was not, or that it was absent when in fact it was present). It is to be used to qualify data. It should never have been used to prevent the reporting of measurements to the laboratory's clients.

The second problem is in the loss of the actual measurement. If the value 0.006 is obtained over several samples it is much more reliable than if obtained once only. Analytical Detection Criterion and Detection Limit are to be used by the analyst to evaluate a single piece of data. In as much as the client can calculate an average, the Detection Limit will be lowered by the square root of the number of data included in the average.

The third difficulty with the traditional approach to "detection limit" is that it gives more credence to the result (eg. 0.011) just above the in-house limit (eg. 0.010) than to the result (eg. 0.009) just below it. Firstly, 0.009 is not significantly different from 0.011, and secondly, even though 0.009 is not high enough to confirm the presence of the constituent, neither is it low enough to deny the presence of as much as perhaps 0.018 units. Thirdly, if the 0.009 is in error so is the 0.011 result.

The fourth difficulty with the "less than" (<) approach to reporting that a measurement is below the "Detection Criterion" or "Detection Limit" or in-house definition of a "Minimum Reportable Value" is that only data above the limit is eligible for inclusion in an average. Therefore such averages will be biased high and the data user will be led to make a Type-I decision error, i.e. concluding the constituent is present when it probably is not.

Summary

Credibility involves presenting both sides of the coin. As analysts we too often try to prove how good our data is. We downplay the possibility of error. We sometimes unwittingly mislead some of our clients by failing to provide complete statements to qualify results in terms of nomenclature, recovery, error and/or reproducibility.

The more routine our work is, the more reliable the data produced becomes, but also the more mechanical. Credibility requires us not only to do things well, but also to do the right thing at the right time. A good quality control program on the bench supported by a formal protocol for documenting success and failure, provides others with assurance that we are doing our job to the best of our ability, and thereby provides us with the credibility we need and desire.

In the final analysis however the data produced by even the most credible laboratory is not better than the validity of the sample submitted for measurement. Quality assurance activity must be extended beyond the laboratory to include program design and field operations. Credibility is enhanced by our ability to solve problems, not by ability to produce numbers.

References

(1) Eisenhart, C.; Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems. Journal of Research of NBS, Vol. 67C, No. 2 (1963).

(2) King, D. E.; Detection of Systematic Error in Routine Trace Analysis, in Proceedings of the 7th Materials Research Symposium, NBS Spec. Pub. 422, Vol. I, p. 141 (Aug. 1976).

(3) King, D. E.; Paired Sample Calibration Control. In prep.

(4) Zar, J. H.; The Power of Statistical Testing: Hypotheses about Means. American Laboratory, 13 (6), 102 (1980).

(5) King, D. E.; Evaluation of Interlaboratory Comparison Data by Linear Regression Analysis, in Proceedings of the 8th Materials Research Symposium, NBS Spec. Pub. 464, p. 581 (Nov. 1977).

(6) Youden, W. J.; Statistical Techniques for Collaborative Tests. AOAC, Washington, D.C. (1967).

(7) Youden, W. J.; Ranking Laboratories by Round-Robin Tests, in NBS Spec. Pub. 300, Vol. I, p. 165 (Feb. 1969).

(8) Aspila, K. et al; Data Quality Assessment. Report of the Data Quality Work Group, Nov. 1980. International Joint Commission, Windsor, Ont., p. 7-8.

(9) King, D. E.; Specification of Analytical Detection Capability. In prep.

D E F I N I T I O N S

CREDIBILITY: FACT OR FANCY

- THE TRUTH AS I PERCEIVE IT!
- LIES IN THE EYE OF THE BEHOLDER?

INCREDIBILITY: CREDIBILITY AT ITS WORST?!

- SELF CONFIDENCE WITHOUT PROOF!

QUALITY CONTROL: DATA AND OBSERVATIONS.

- ARE THINGS O.K. TODAY?
- TASKS PERFORMED BY TECHNICAL STAFF.

QUALITY ASSURANCE: PROTOCOLS AND EVALUATIONS

- HOW WERE THINGS LAST MONTH?
- TASKS PERFORMED BY SUPERVISORY STAFF.

CREDIBILITY: THE NATURAL OUTCOME OF

- A WELL DESIGNED QC PROTOCOL IMPLEMENTED DAILY
- DATA RECORDED AND ORGANIZED FOR REGULAR RETRIEVAL
- DOCUMENTED PROCEDURES FOR QA
- DOCUMENTED QA EVALUATIONS

MEASUREMENT

PRECISION:

- ABILITY OF A METHOD TO PRODUCE THE SAME ANSWER WITHIN A NARROW RANGE.

ACCURACY:

- ABILITY OF A METHOD TO OBTAIN THE 'CORRECT' ANSWER AS DEFINED BY AN INTERNATIONAL STANDARD.

REPEATABILITY:

- ABILITY OF AN ANALYST TO PRODUCE THE SAME ANSWER, WITHIN A NARROW RANGE, UNDER THE SAME CALIBRATION CONDITIONS.

REPRODUCIBILITY:

- ABILITY OF AN ANALYST TO PRODUCE THE SAME ANSWER, WITHIN A NARROW RANGE, UNDER DIFFERENT CALIBRATION CONDITIONS.

BIAS:

- DIFFERENCE BETWEEN ANSWERS AS PRODUCED BY DIFFERENT ANALYSTS USING THE SAME OR DIFFERENT METHODS.

ERROR:

- DIFFERENCE BETWEEN THE AVERAGE ANSWER AND THE 'CORRECT' VALUE AS DEFINED BY AN INTERNATIONAL STANDARD.

CREDIBILITY FACTORS TO BE CONSIDERED WHEN IMPLEMENTING A FIELD PROGRAM

1. RATIONALE
   - WHY ARE WE OUT THERE?

2. RESOURCES
   - WHAT SPECIAL FIELD OR LAB CAPABILITIES ARE NEEDED? DO WE HAVE THEM?

3. SAMPLING DESIGN
   - ARE WE SAMPLING THE RIGHT SPOTS?

4. FIELD ACTIVITIES
   - WHAT HAS TO BE DONE IN THE FIELD TO MAKE THE SAMPLES WORTH THE COST OF ANALYSIS?

5. SAMPLE IDENTITY AND INTEGRITY
   - ARE THE SAMPLES CORRECTLY AND SUFFICIENTLY IDENTIFIED?

6. DATA INTERPRETATION
   - WILL OUR CONCLUSIONS BE SUPPORTED BY THE DATA PRESENTED?
   - HOW GOOD IS THE DATA?

7. LABORATORY CREDIBILITY.

CREDIBILITY FACTORS CONSIDERED IN THE LABORATORY

1. PHYSICAL STRUCTURE
   - CLEANLINESS, SAFETY, SPACE
   - AIR CONDITIONING, DUST CONTROL (TEMPERATURE, HUMIDITY, ETC.)
   - STORAGE FACILITIES

2. MECHANICAL CONDITION OF EQUIPMENT
   - MAINTENANCE PROGRAM
   - SAFELY CONNECTED.

3. SOURCE AND QUALITY OF REAGENTS
   - SAFETY AND STORAGE
   - DISTILLED WATER AND SOLVENTS

4. DOCUMENTED ANALYTICAL PROCEDURES
   - RELATED TO SAMPLE TYPES BEING ANALYSED
   - SUITABLY PRECISE AND ACCURATE FOR MEETING CLIENT NEEDS.

5. DOCUMENTED QA AND QC PROTOCOLS.

6. ANALYST EXPERIENCE AND PROFICIENCY
   - TRAINED TO RECOGNIZE SPECIAL CASES.

7. CALIBRATION MATERIALS AND PROCESSES
   - SOURCE AND RELIABILITY
   - PREPARATION AND MAINTENANCE
   - EXTERNAL REFERENCES

QUALITY ASSURANCE

DOCUMENTATION

1. DAILY QUALITY CONTROL PROTOCOL

   - FACTORS TO BE CONTROLLED, LIMITS
   - RECORDS TO BE MAINTAINED
   - PROCEDURE FOR VERIFYING CONTROL
   - ACTION TO BE TAKEN IF OUT-OF-CONTROL

2. QUALITY ASSURANCE PROCEDURES

   - DATA EVALUATION TECHNIQUES
   - LONG-TERM ASSESSMENT

3. SUPERVISORY RESPONSIBILITIES

   - DAILY VERSUS LONG-TERM
   - FREQUENCY OF QC DATA EVALUATION
   - ACTION TO BE TAKEN TO DETERMINE PERFORMANCE AND IMPROVE IT.

4. SUMMARIES OF DAILY QC SUCCESS

   - HOW WELL WAS THE SYSTEM RUN
   - FREQUENCY OF FAILURE, ACTION TAKEN
   - SIGNIFICANCE TO CLIENT

5. DATA COMPARABILITY

   - FREQUENCY OF AND WITH WHOM WERE EXTERNAL COMPARISONS MADE.

6. HISTORICAL RECORDS.

   - PREVIOUS PROCEDURES, RELIABILITY
   - DEVELOPMENT OF CURRENT METHOD
   - SUITABILITY TO SAMPLE TYPES BEING ANALYSED.

QUALITY CONTROL

1. CONTROL MUST BE ESTABLISHED BEFORE IT CAN BE MAINTAINED!

2. A SYSTEM IS OUT-OF-CONTROL IF IT PRODUCES UNEXPECTED DATA MORE THAN ONCE IN TWENTY TO TWENTY-FIVE RUNS!

3. CONTROL LIMITS TEND TO TIGHTEN WITH TIME AS CONTROL PROTOCOL IS ESTABLISHED.

4. ACCEPTANCE LIMITS, LOOSER THAN CONTROL LIMITS, PERMIT RELEASE OF DATA TO CLIENTS WHO DO NOT REQUIRE PERFORMANCE EQUAL TO THE SYSTEM CAPABILITY.

5. BOTH COMPONENT AND PRODUCT CONTROLS ARE REQUIRED TO MINIMIZE SYSTEM FAILURE. PREVENTION RATHER THAN CURE!

6. CONTROL CHARTING PERMITS TREND ANALYSIS!

7. DON'T FIDDLE OR TRENDS, WHICH COULD IDENTIFY CAUSE, WILL BE DESTROYED!

8. CONTROL IS REQUIRED OVER

   A) REPEATABILITY

   B) REPRODUCIBILITY

   C) BIAS/ERROR/RECOVERY